Supporting Information

IsRNA: An iterative simulated reference state approach to modeling correlated interactions in RNA folding

Dong ZHANG and Shi-Jie CHEN[*]

Department of Physics, Department of Biochemistry, and MU Informatics Institute

University of Missouri, Columbia, MO 65211

Details of coarse-grained force field

The corresponding energy functions for the three local energy terms are assumed to have the following functional form:

$$E_{\text{bond}}(b) = K_b(b-b_0)^2 + \frac{H_b}{\sqrt{2\pi}\sigma_b}e^{-\frac{(b-b_1)^2}{2\sigma_b^2}} \tag{1}$$

$$E_{\text{angle}}(\theta) = K_a(\theta-\theta_0)^2 + \frac{H_a}{\sqrt{2\pi}\sigma_a}e^{-\frac{(\theta-\theta_1)^2}{2\sigma_a^2}} \tag{2}$$

$$E_{\text{torsion}}(\phi) = \sum_{n=1}^{4} K_n[1+\cos(n\phi-\phi_n)] \tag{3}$$

where b, θ and ϕ denote the bond length between two connected beads, the bond angle between two adjacent bonds, and the torsion angle between three successive bonds, respectively. Parameters K_b, b_0, H_b, σ_b, and b_1 in eq 1 describe the strength of bond stretching and the equilibrium bond length. Parameters K_a, θ_0, H_a, σ_a and θ_1 in eq 2 characterize to the bending energy of the bond angle and the equilibrium bond angle. K_n and ϕ_n ($n = 1, 2, 3, 4$) in eq 3 are related to the local minima of the torsional energy. The above parameters will be determined using the frequency distributions of the respective structure parameters of the known RNA structures. Compared with the widely used harmonic potentials for bond stretching and bond angle terms, extra Gaussian terms are introduced here in order to enable broader sampling of the backbone conformations and to capture more precise details of the statistical potential.

The non-local term E_{pair} accounts for base-base, including base pairing and stacking, and base-backbone interactions:

$$E_{\text{pair}}(r) = D_0[e^{-2\alpha(r-r_0)} - 2e^{-\alpha(r-r_0)}] + \frac{H_1}{\sqrt{2\pi}\sigma_1}e^{-\frac{(r-r_1)^2}{2\sigma_1^2}} + \frac{H_2}{\sqrt{2\pi}\sigma_2}e^{-\frac{(r-r_2)^2}{2\sigma_2^2}}, r < r_{cut} \tag{4}$$

Here r is the pairwise distance between two non-neighboring coarse-grained beads. The first Morse potential term in eq 4 is designed to give the overall shape of the knowledge-based statistics/energies while the later two Gaussian terms account for local features of the statistical distribution. All the related parameters D_0, α, r_0, H_1, σ_1, r_1, H_2, σ_2, r_2, and the cut-off distance r_{cut} for the different types of pairwise interactions will be determined from the iterative simulated procedure described below.

The electrostatic term E_{ele} accounts for backbone-backbone electrostatic interactions:

$$E_{\text{ele}}(r) = \frac{Cq_iq_j}{\epsilon_r r}e^{-r/\xi} + \frac{H_e}{\sqrt{2\pi}\sigma_e}e^{-\frac{(r-r_e)^2}{2\sigma_e^2}}, r < r_{cut} \tag{5}$$

Here $\epsilon_r = 78.285$ is the dielectric constant of the solvent at room temperature and each phosphate group is assumed to carry one unit of the electron charge. We set the Debye length to be $\xi = 10$Å (for 100 mM Na^+). An additional Gaussian term is also introduced in eq 5 to account for the local details of the knowledge-based potential. Parameters H_e, σ_e and r_e will be determined below. We choose $r_{cut} = 12.5$Å as the cut-off distance and set $E_{\text{ele}}(r) = 0$ for pairwise distance $r > r_{cut}$.

The last term E_{LJ} describes the excluded volume interaction between any two non-connected coarse-grained beads through the repulsive Lennard-Jones potential

$$E_{\text{LJ}}(r) = \epsilon[(\sigma/r)^{12} - (\sigma/r)^6], r < \sigma \tag{6}$$

Here the energy constant is set to be $\epsilon = 0.5$ kcal/mol, and $\sigma = (\sigma_i + \sigma_j)/2$ is for the interaction between the i-th and the j-th types of the coarse-grained beads with diameters σ_i and σ_j, respectively.

Technical details for coarse-grained MD simulation

For a given coarse-grained force field, the coarse-grained molecular dynamics (CGMD) simulation starts out from a native (PDB) structure with coarse-grained RNA representation. The simulation implements Langevin dynamics (NVT ensemble) in the modified open source software, LAMMPS. The time step for integration was set to $\Delta t = 1$ fs and the simulation temperature was T =298 K. After the first 4×10^7 equilibration steps (40 ns), an additional 2×10^7 simulation steps (20 ns) were used to collect the snapshots at an interval of 10^4 steps (10 ps). In total, CGMD simulation generates a conformational ensemble of 8×10^4 structures for construction of the reference state.

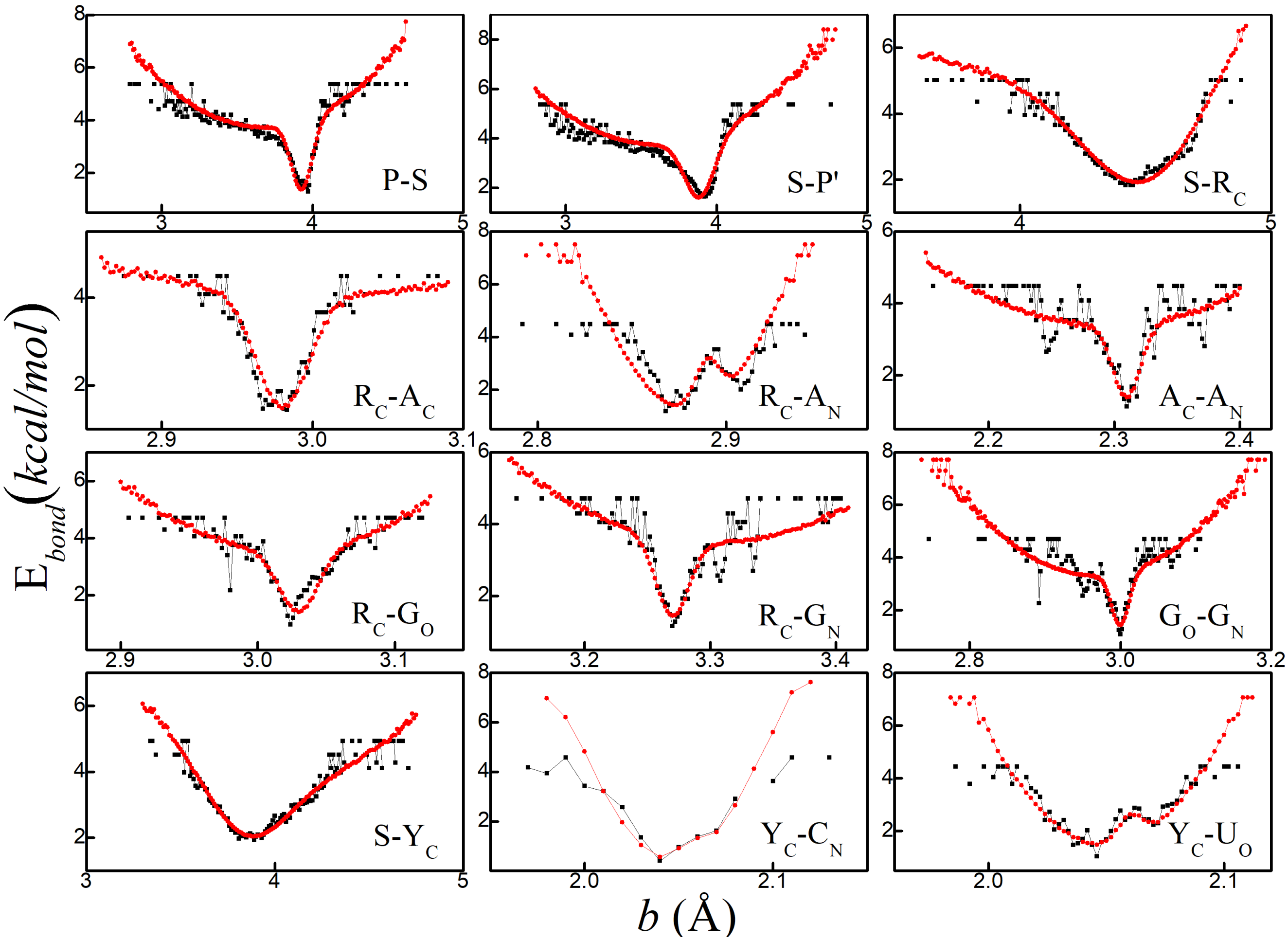


Figure 1: An overview of the 12 types of extracted bond stretching energies in IsRNA. The black square is the observed statistical distribution ($E_{obs}(b)$) derived from the PDB structures, the red circle is the simulated Boltzmann-like distribution ($E_{sim}(b)$).

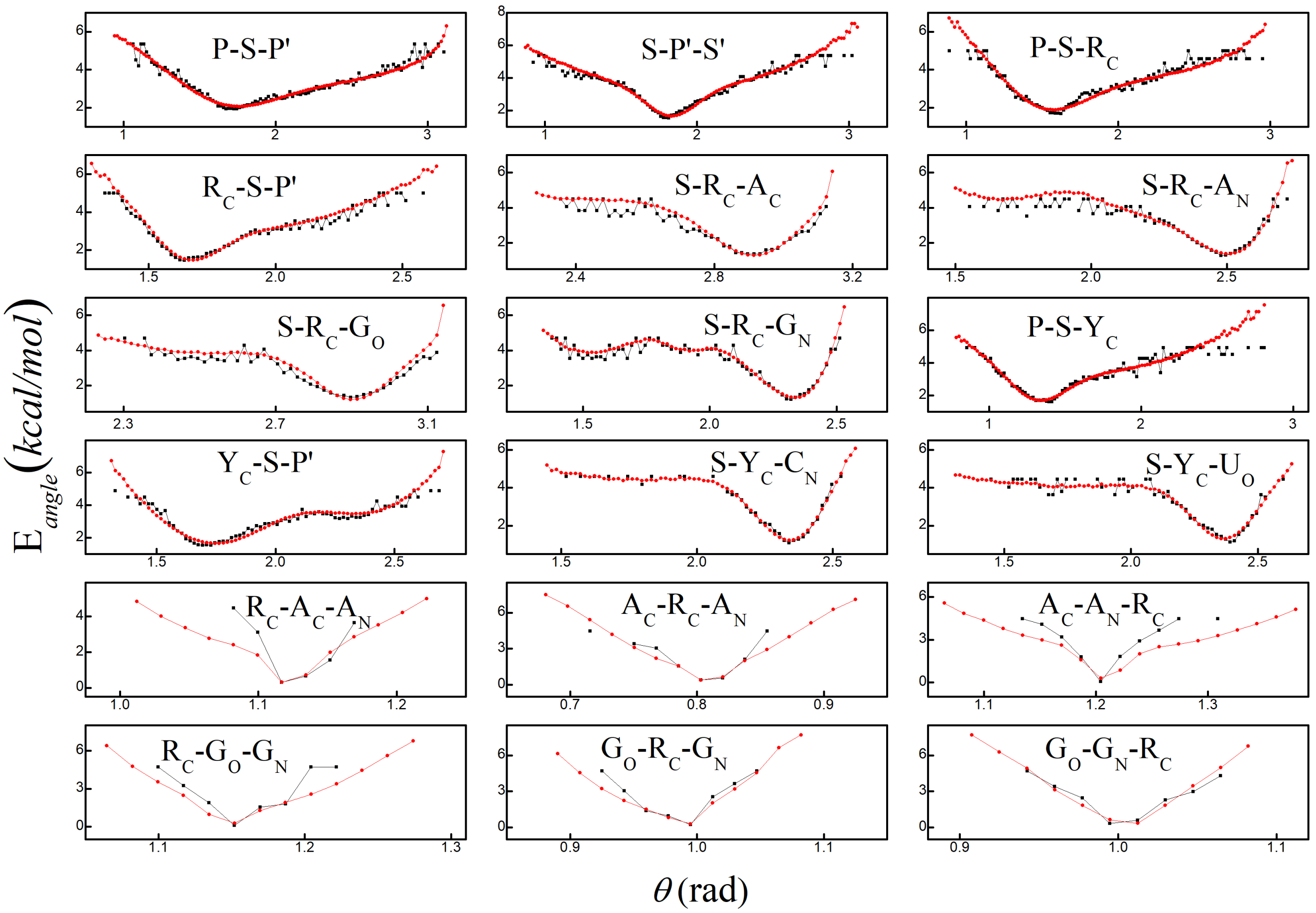


Figure 2: An overview of the 18 types of bond angle bending energies in IsRNA. The black square is the distribution ($E_{obs}(\theta)$) derived from the PDB structures, the red circle is the simulated Boltzmann-like distribution ($E_{sim}(\theta)$). Only the first 12 terms are parameterized, leaving the later 6 terms free.

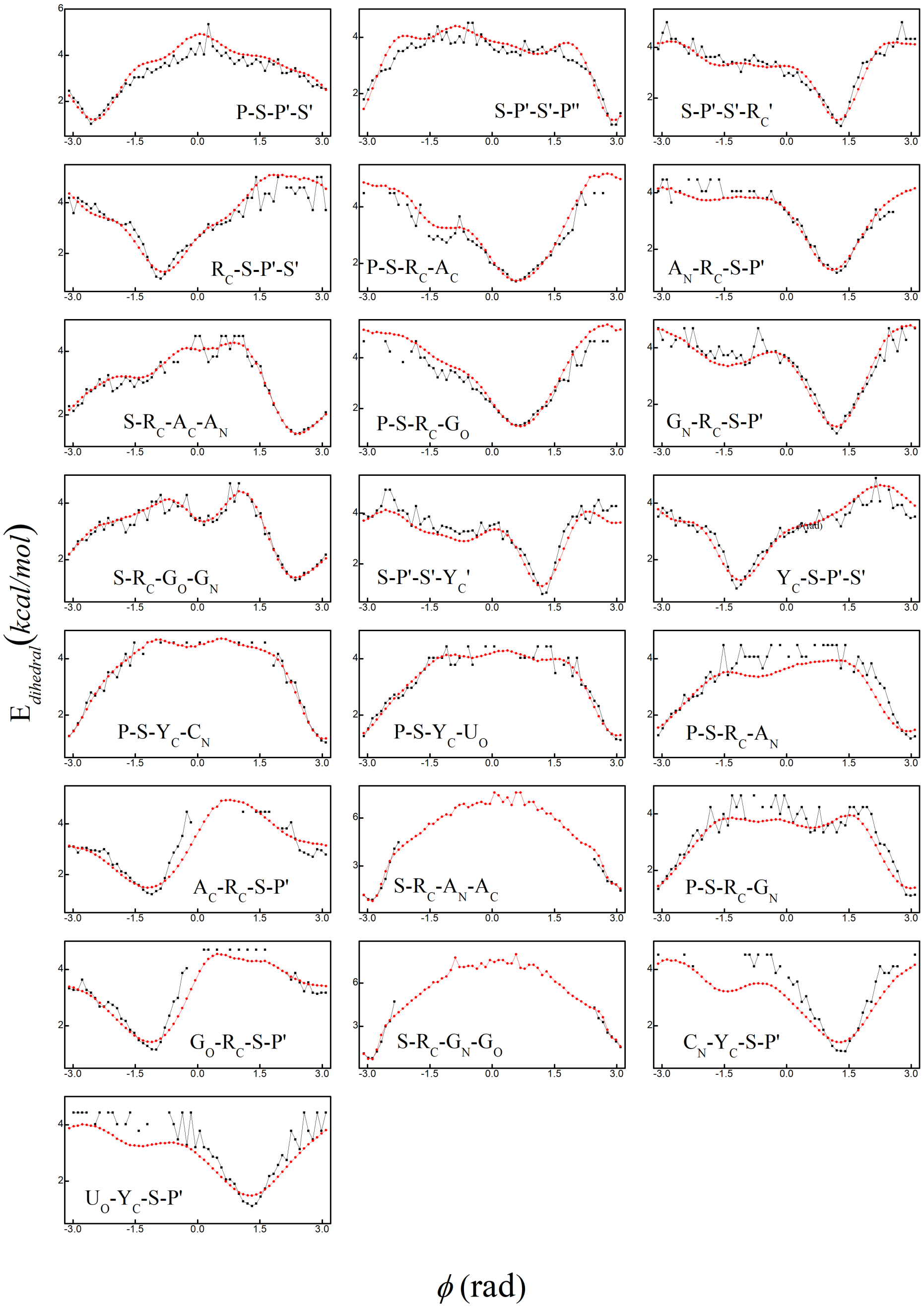
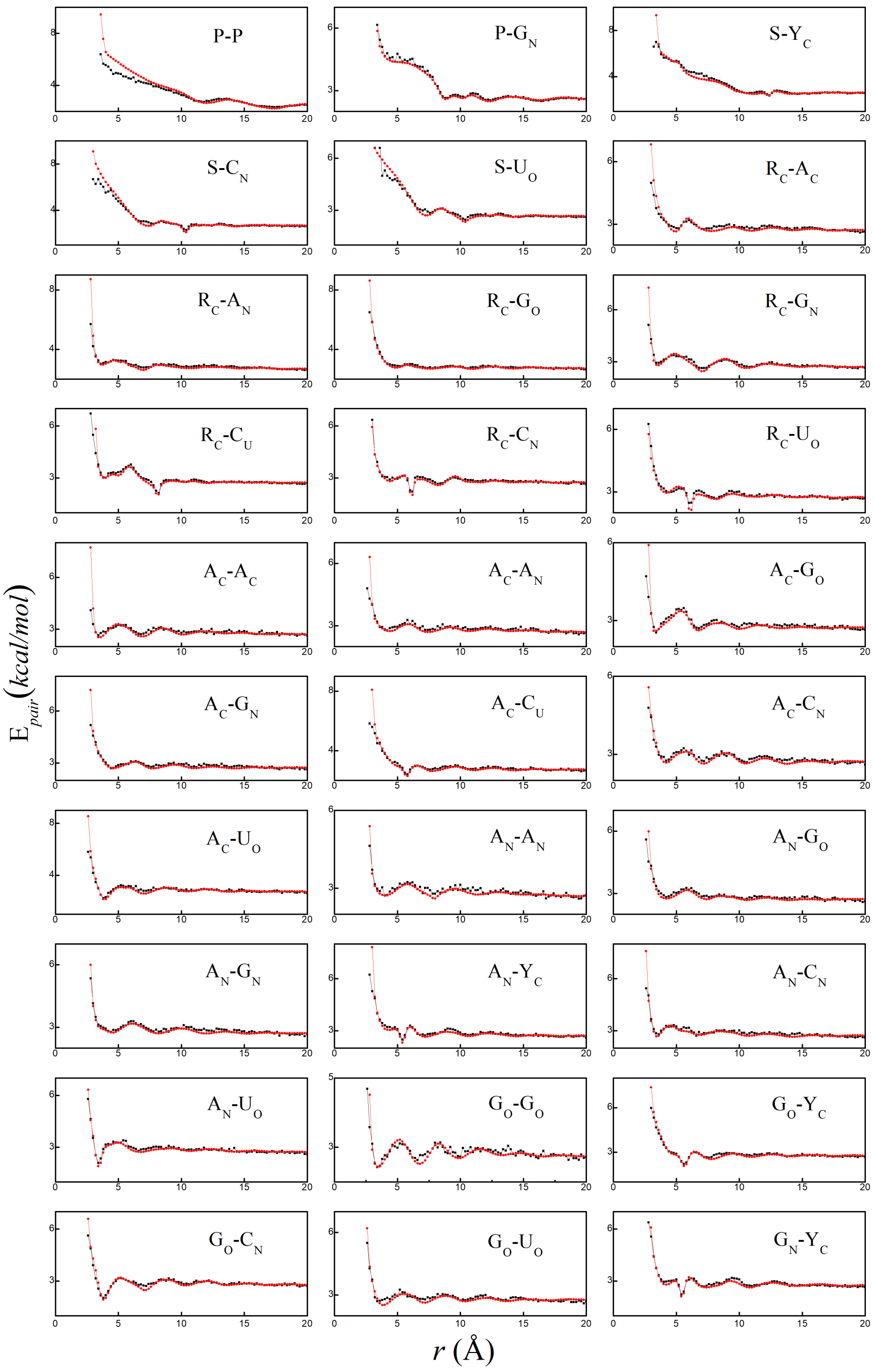


Figure 3: An overview of the 22 types of torsion energies in IsRNA. The black square is the distribution ($E_{obs}(\phi)$) derived from the PDB structures, the red circle is the simulated Boltzmann-like distribution ($E_{sim}(\phi)$). Only the first 14 terms are parameterized, leaving the later 8 terms free.


P-P
P-G_N
S-Y_C
S-C_N
S-U_O
R_C-A_C
R_C-A_N
R_C-G_O
R_C-G_N
R_C-C_U
R_C-C_N
R_C-U_O
A_C-A_C
A_C-A_N
A_C-G_O
A_C-G_N
A_C-C_U
A_C-C_N
A_C-U_O
A_N-A_N
A_N-G_O
A_N-G_N
A_N-Y_C
A_N-C_N
A_N-U_O
G_O-G_O
G_O-Y_C
G_O-C_N
G_O-U_O
G_N-Y_C
E_pair (kcal/mol)
r (Å)
0
5
10
15
20

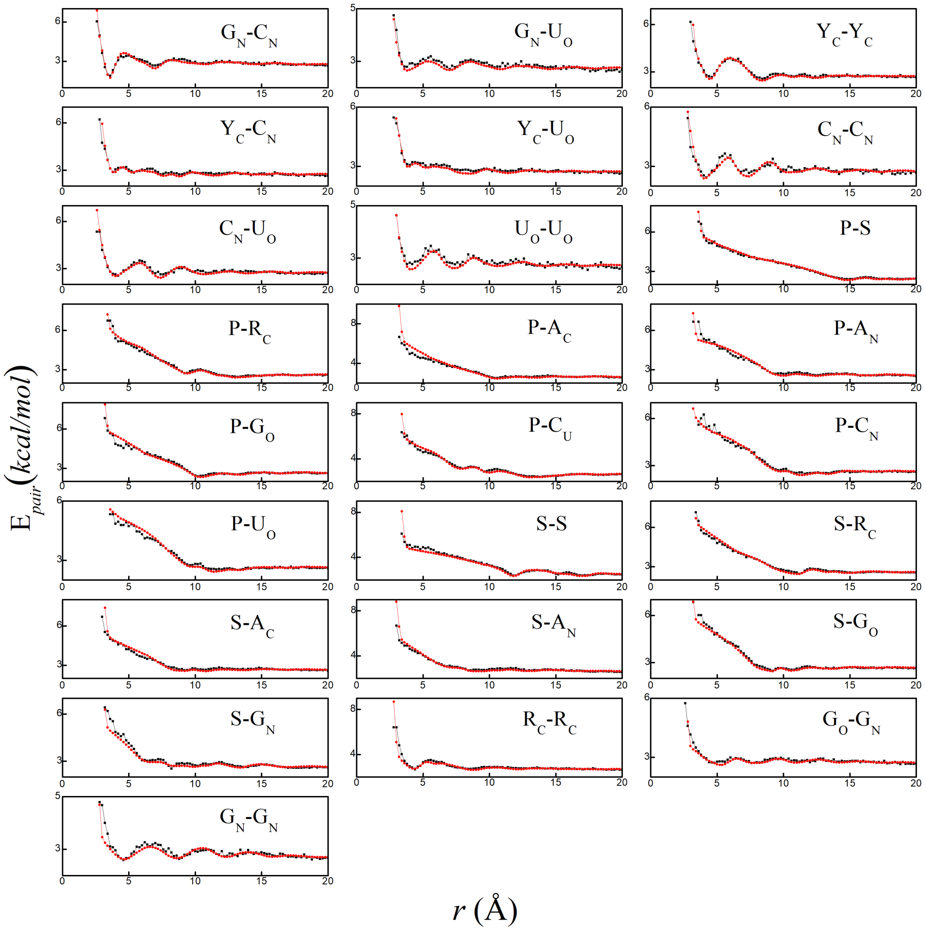


Figure 4: An overview of the 55 types of non-local pairwise interactions in IsRNA. The black square is the distribution ($E_{obs}(r)$) derived from the PDB structures, the red circle is the simulated Boltzmann-like distribution ($E_{sim}(r)$). Only the first 38 terms are parameterized, leaving the other later 17 terms free.

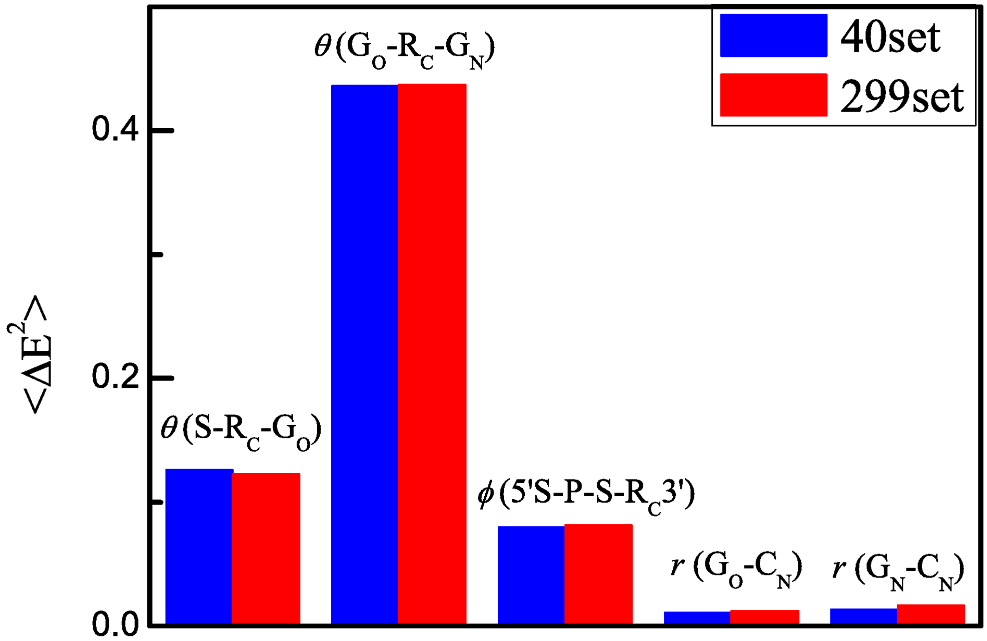


Figure 5: Results for the extra benchmark over the whole data set of 299 structures. The average squared errors between the observed and simulated probabilities ($< \Delta E^2 >= \sum_{i=0}^{n} |E^i_{\rm obs} - E^i_{\rm sim}|^2/n$, n is the number of discrete data) are given for five illustrated structure parameters in the main text. Results for "40set" are from the fixed dataset of 40 structures (training set in the main text). Results for "299set" are from the coarsed grained MD simulations over the whole dataset of 299 structures using the energy parameters obtained from the fixed training set.

Table 1: Properties of ten types of CG beads

CG bead	Mass (amu)	Diameter (Å)	Grouped heavy-atom
P	94.97	4.0	P, OP1, OP2, O3, O5’
S	92.05	3.8	C5’, C4’, O4’, C3’, C2’, O2’, C1’
$\mathrm{R_C}$	78.05	3.2	N9, C8, N7, C5, C4, N3
$\mathrm{A_C}$	26.02	3.0	C6, N6
$\mathrm{A_N}$	26.02	3.0	N1, C2
$\mathrm{G_O}$	28.01	3.0	C6, O6
$\mathrm{G_N}$	40.03	3.0	N1, C2, N2
$\mathrm{Y_C}$	38.03	3.4	N1, C5, C6
$\mathrm{C_N}$	68.04	2.6	C2, O2, N3, C4, N4
$\mathrm{U_O}$	70.03	2.6	C2, O2, N3, C4, O4

Table 2: PDB ids of 40 simulated structures for iterative simulated construction of the reference states.

Simulated dataset							
1a4d	1bau	1c2x	1cq5	1e95	1f5u	1kaj	1kpd
1p5p	1rmn	1rnk	1t28	1txs	1ymo	1z2j	1z43
1zo3	2d1a	2d1b	2f4x	2g1w	2gm0	2jyj	2k4c
2k95	2ke6	2kur	2kyd	2l1f	2l2j	2lkr	2m23
2mhi	2n1q	2p89	2tpk	3a3a	3d0x	4p8z	299d